AIntelia

An Outlier Analysis on Multi-Dimensional and Time-Series Data

Ümmügülsüm MENGUTAYCI[1,3*], Selma Ayşe ÖZEL[2]
[1]*Cukurova University Institute Of Natural and Applied Sciences, Adana, Turkey, mengutayci@tarsus.edu.tr, ORCID: 0000-0001-9861-8957*
[2]*Cukurova University, Adana, Turkey, saozel@cu.edu.tr, ORCID: 0000-0001-9201-6349*
[3*]*Tarsus University, Mersin, Turkey, mengutayci@tarsus.edu.tr, ORCID: 0000-0001-9861-8957*

Outlier detection refers to the detection of unexpected situations in the data. Outliers are fraud, hacking, mislabeled data, or unusual behavior in the system. Therefore, it is important to determine these values. In this study, outlier detection performances of the algorithms used in outlier detection analysis on different types of data sets were calculated and compared. As a result of the study, it was seen that the algorithms showed sufficient success. The highest performance was seen in the Histogram-based outlier detection algorithm with 99 % accuracy.

Keywords: *Outlier detection, multivariate, time-series, dataset.*

1. Introduction

Outlier detection is used in many applications in data mining[1]. Outliers are a data object that deviates significantly from other objects in the data set, as shown in Figure 1 [2]. In Figure 1, the R region doesn't follow the same distribution as other objects in the data set, so the R region can be defined as an outlier. Outliers occur as a result of data entry errors caused by changes in system behavior, fraud, forgery behavior, installation error, human errors during data collection [1]. Determining these values is of great importance to ensure the security of the system and to obtain more accurate results. Detection of outliers is of great importance for many applications such as fraud detection, intrusion detection, public safety, healthcare, damage detection, image processing [2].
Outliers are often in the minority in the data set. Therefore, it can be difficult to detect. Besides the scarcity of their numbers, as the size of the dataset increases, the data becomes more sparse and it can be difficult to capture neighborhood information because it becomes difficult to estimate the distances and density between the data [3]. In this study, experiments were carried out on different data sets to analyze the efficiency of anomaly detection on multidimensional data. In addition, the anomaly detection performance of algorithms on time series [4], which is widely used in military, economic and scientific fields, is examined on a multidimensional time series data set.
In this study, CBLOF-BIRCH, LOF, k nearest neighbor, Angle-Based, histogram-based outlier detection algorithms in which Isolation Forest, CBLOF, CBLOF and BIRCH algorithms are used together are used to detect outliers on breast-cancer, pendigits and SKAB datasets. When the literature is examined, these data sets and algorithms have been used in many different ways in outlier detection. When the anomaly detection studies on breast-cancer and pendigits datasets are examined, it is seen that the studies generally focus on the LOF and Isolation Forest algorithms [5] [6] [7] [8] [9]. When the anomaly studies in the literature related to the SKAB dataset are examined, mostly LSTM, AutoEncoder, convolutional neural network, RNN, CPDE, MSET, Isolation Forest methods are used. [10] [11] [12][13] [14]. LOF, Isolation Forest, KNN, CBLOF algorithms have been frequently used in the literature for anomaly detection in different application areas such as medical applications and the banking sector. [5] [9] [15] [16] [17]. In addition, it has been stated in the literature [18] that the LOF algorithm gives successful results in intrusion detection. Anomaly detection studies on different time series data sets of the histogram-based anomaly method were encountered [15] [19] [20].

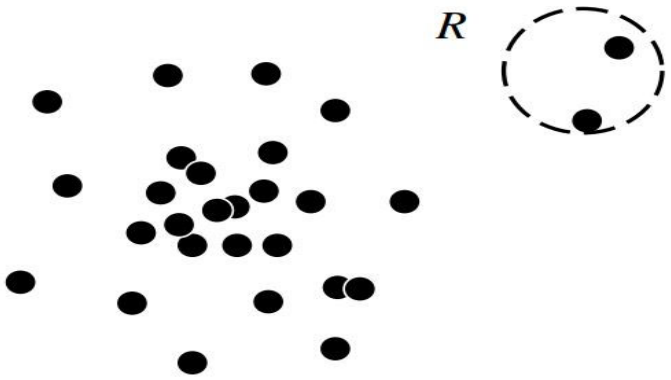


Figure 1. *R is outlier* [2].

2. Related Work

Some studies have been carried out in the literature on the detection of outliers. Some of them are mentioned in this section.
In [23], outliers were detected using probabilistic, proximity-based and linear models in IoT-based structural health monitoring systems. For this, they used the shuttle data set in the UCI database.
In this study [24], to find outliers on diabetes data obtained from a hospital, random forest, KNN, support vector machine algorithms and their proposed method hierarchical clustering-based support vector machine (HCSVM) method were used to analyze the performance of these algorithms. The results showed that the HCSVM algorithm performed the best outlier detection in the data set with 4805 normal data and 96 abnormal data.
In [25], a new method, C-LSTM, consisting of a combination of convolutional neural network (CNN) and long short-term memory (LSTM) algorithms, was developed to detect abnormal values on a web traffic dataset Webscope S5 with a one-dimensional time series signal. With the developed method, outliers in the dataset were detected.
[26] used logistic regression, decision tree, k-nearest neighbor, random forest and autoencoder methods to detect fraud behavior in credit card transactions.
In another study [27], fraud detection in online games and games of chance was investigated with clustering-based algorithms.
In this study, multivariate and time series data sets of different sizes were used. The success of outlier detection on multidimensional data of Angle-based algorithms, which are claimed to have good performance in high-dimensional data [21], Isolation Forest, KNN, CBLOF, LOF, Histogram based algorithms, which are frequently used in different outlier detection applications in the literature, and BIRCH algorithm (was used with CBLOF), which is a good clustering method for large databases [22], algorithms has been examined.

3. Methodology

A. Dataset:

In the study, breast-cancer [6], which consists of digitized features of a breast mass, pendigits [8] consisting of handwritten samples, and SKAB [10], which consists of time series data developed for anomaly studies, were used.
The datasets were divided into 80 percent training and 20 percent test data. Table 1 shows the number of data samples, data size, data types, number and percentages values of outliers in each dataset. Also, in Table 1, the duration column shows the time period that the dataset contains.

Dataset Name	Number of instances	Number of dimensional	Attribute type:	Number of outliers	Percentage of outliers	duration
Breast-cancer	366	30	Real	10	2.73 %	-
pendigits	6870	16	Numeric	156	2.27 %	-
SKAB	22473	10	Categorical, Numeric	7826	34.8 %	1 day (2020-03-09)

Table 1. *Data properties.*

B. Methods:

Angle-based Outlier Detection (ABOD):

It has been developed for high-dimensional data. The angle of each of them in the dataset with all the other data is looked at. The variance of each angle is calculated. If the result is less than the predetermined value, this data is considered outlier [21].

Local Outlier Factor (LOF):

For each object in the dataset, a local outlier factor is determined, representing the outlier value. The LOF of most objects in a cluster is equal to 1. Min and max values are determined for other objects. The density distance of each object from its neighbors is measured. Data with a lower density than its neighbors is considered outlier. This method is related to the density-based clustering method [28].

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH):

It is a hierarchical clustering method. It has two features, Clustering Feature and Clustering Feature tree. A Clustering Feature Tree is initially created for clustering. Then, the entire dataset is scanned and the clustering feature value is calculated by creating subsets with the predetermined N value [2].

Cluster-based Local Outlier Factor* (*CBLOF):

Computes an outlier score based on the cluster-based local outlier factor [29]. In cluster-based outlier detection, anomaly data occurs in three ways. The data may not belong to any cluster, the distance between the data and the closest cluster may be too far, or the outlier data may be a sparse cluster [2]. The CBLOF method calculates the size of the cluster to which the data belongs and the distance of this data from the cluster center to find outliers [30].

Histogram Based Outlier Detection (HBOS):

Unsupervised method that calculates outlier degree by generating histograms [29]. It uses histograms to detect normal and outlier data. The height of the created histogram is of great importance in determining the anomaly data correctly. Because if the size is small, the accuracy of the normal data being at the height in the specified range decreases. Conversely, if it is large, it can cause outliers to appear as normal data. It is widely used especially in fraud detection applications [30].

K-Nearest Neighbors (KNN):

The outlier score is calculated by taking the k-nearest neighbor of each data. It uses neighborhood information to detect outliers [1].

Isolation Forest:

It randomly selects a feature and isolates observations by randomly choosing a split value between the minimum and maximum values based on the selected feature [9].

C. Tools:

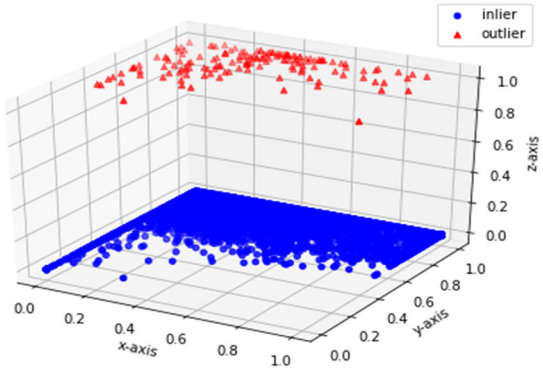
In this research, the Python Pyod module was used to detect outliers.

D. Evaluation:

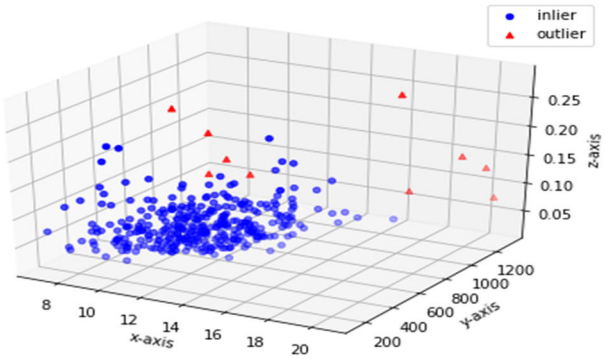
ROC analysis was developed to examine the performance states of the systems. ROC charts are used to show classification performances. In the ROC graph, the x-axis represents the false positive rate, that is, the proportion of misclassified data in the dataset, and the y-axis represents the proportion of correctly classified data in the dataset, the true positive rate. The area under a ROC curve is defined by the AUC. The larger this area, the better the success of the algorithm is considered [31]. The ROC-AUC value was taken as a criterion to analyze the success of the algorithms used in this study.

E. Results:

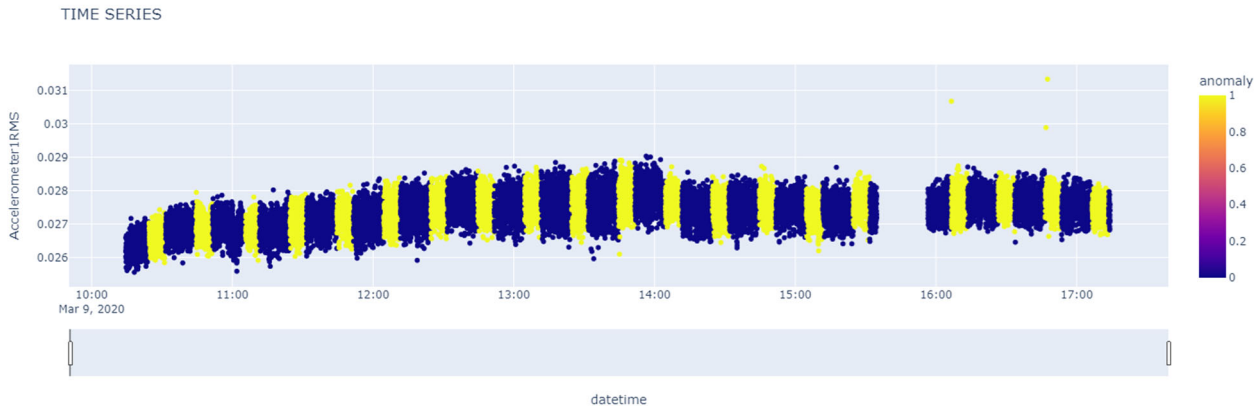
Abnormal and normal data samples in the datasets used in this study are shown in Figure 2 in three dimensions. When we look at the pendigits and breast-cancer datasets, it is seen that the abnormal data stand further from the normal data distribution. In the time series SKAB dataset, on the other hand, the outliers took place in the usual time period.



(a) pendigits dataset

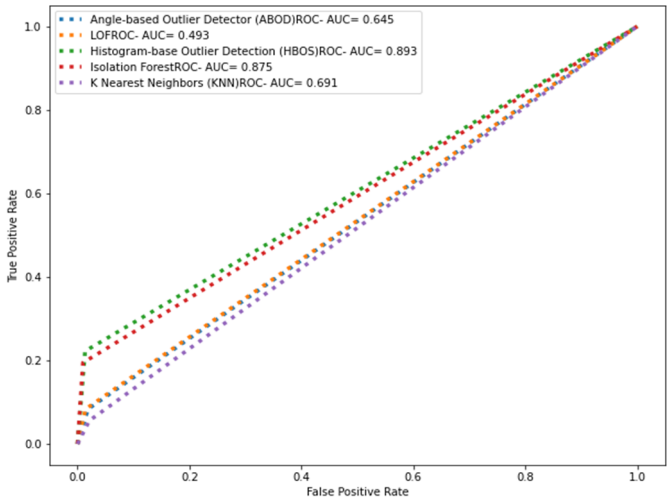


(b) breast-cancer dataset

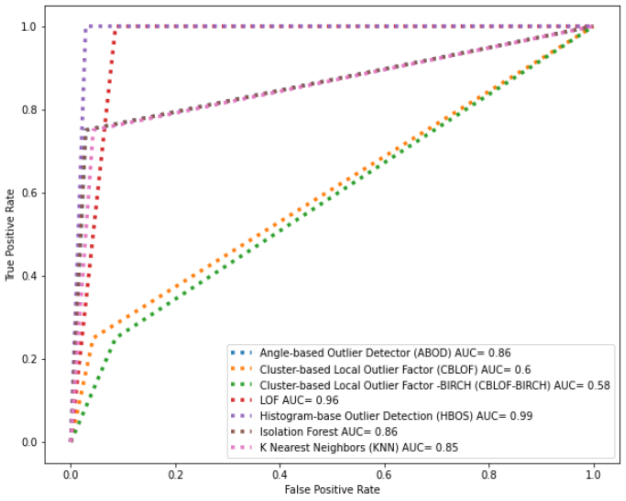


(c)Time-dependent variation of the feature determined in the SKAB dataset. Outliers are represented by yellow points.

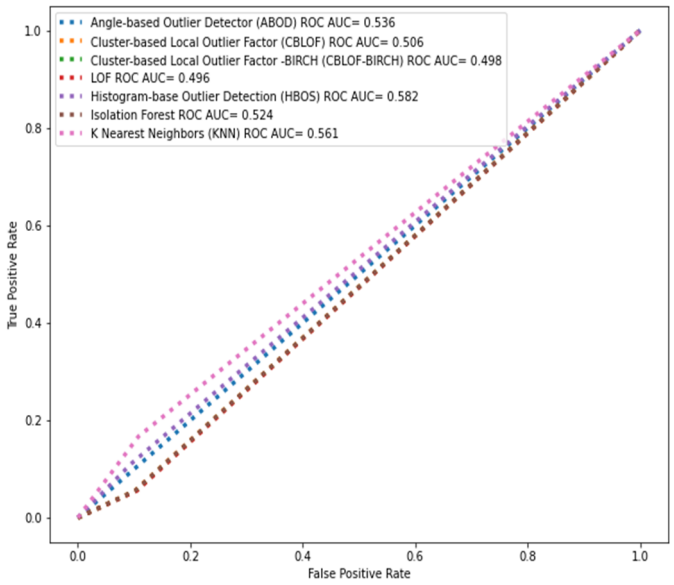
Figure 2. *The distribution of outlier and normal data in datasets.*



(a) pendigits dataset



(b) breast-cancer dataset



(c) SKAB dataset

***Figure 3**. The AUC distribution of algorithms applied to different datasets.*

The performance distributions of each algorithm were shown in Figure 3 for the SKAB, breast-cancer, pendigits datasets by using ROC-AUC analysis graphs. As a result of the experiments, the HBOS algorithm performed better than other algorithms in detecting outliers in all data sets used. The HBOS algorithm showed its best performance on the breast-cancer dataset with a 99 % success rate. The lowest distribution on the datasets was seen on the SKAB dataset, which is a multivariate time series dataset. When the experimental results are examined in detail, it is seen that although the LOF algorithm has a high success rate of 96 % on the breast-cancer dataset, the same algorithm contains the lowest performance value of the study on the other two datasets, pendigits and SKAB datasets. All results of the study are presented in Table 1.

Algorithm	pendigits	breast-cancer	SKAB
ABOD	0.64	0.86	0.53
CBLOF	-	0.6	0.50
CBLOF-BIRCH	-	0.58	0.498
LOF	0.49	0.96	0.496
HBOS	**0.89**	**0.99**	**0.58**
Isolation Forest	0.87	0.86	0.52
KNN	0.69	0.85	0.56

Table 2. *The success of the algorithms on different datasets. The performance values shown represent ROC-AUC values.*

4. Conclusion

With the widespread use of the Internet, outlier detection has become even more important in order to prevent situations such as forgery and fraud. In this study, the performance of outlier detection algorithms on different types of data sets is compared. As a result of the study, the histogram-based algorithm showed the highest success on the datasets used in the study. In the study, the algorithms generally have a lower success rate on the SKAB dataset than other datasets. In future studies, outlier analysis will be carried out on image data.

REFERENCES

[1] H. Wang, M. J. Bah, and M. Hammad, "Progress in Outlier Detection Techniques: A Survey," *IEEE Access*, vol. 7, pp. 107964–108000, 2019, doi: 10.1109/ACCESS.2019.2932769.

[2] J. Han, M. Kamber, and J. Pei, "Data Preprocessing," *Data Min.*, pp. 83–124, Jan. 2012, doi: 10.1016/B978-0-12-381479-1.00003-4.

[3] H. Liu, X. Li, J. Li, and S. Zhang, "Efficient Outlier Detection for High-Dimensional Data," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 48, no. 12, pp. 2451–2461, Dec. 2018, doi: 10.1109/TSMC.2017.2718220.

[4] H. S. Wu, "A survey of research on anomaly detection for time series," *2016 13th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. ICCWAMTIP 2017*, pp. 426–431, Oct. 2017, doi: 10.1109/ICCWAMTIP.2016.8079887.

[5] D. Xu, Y. Wang, Y. Meng, and Z. Zhang, "An improved data anomaly detection method based on isolation forest," *Proc. - 2017 10th Int. Symp. Comput. Intell. Des. Isc. 2017*, vol. 2, pp. 287–291, Feb. 2018, doi: 10.1109/ISCID.2017.202.

[6] M. Goldstein, "Unsupervised Anomaly Detection Benchmark," 2015.

[7] S. Behera and R. Rani, "Comparative analysis of density based outlier detection techniques on breast cancer data using hadoop and map reduce," *Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016*, vol. 2, Jul. 2016, doi: 10.1109/INVENTIVE.2016.7824883.

[8] F. Keller, E. Müller, and K. Böhm, "HiCS: High Contrast Subspaces for Density-Based Outlier Ranking".

[9] F. Tony Liu, K. Ming Ting, and Z.-H. Zhou, "Isolation Forest".

[10] I. D. K. and V. O. Kozitsin, "Skoltech Anomaly Benchmark (SKAB).," *Kaggle*, 2020. https://www.kaggle.com/dsv/1693952 (accessed May 24, 2022).

[11] I. Katser, V. Kozitsin, V. Lobachev, and I. Maksimov, "Unsupervised Offline Changepoint Detection Ensembles," *Appl. Sci. 2021, Vol. 11, Page 4280*, vol. 11, no. 9, p. 4280, May 2021, doi: 10.3390/APP11094280.

[12] A. Garg, W. Zhang, J. Samaran, R. Savitha, and C. S. Foo, "An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series," *IEEE Trans. Neural Networks Learn. Syst.*, 2021, doi: 10.1109/TNNLS.2021.3105827.

[13] X. Wang, D. Pi, X. Zhang, H. Liu, and C. Guo, "Variational transformer-based anomaly detection approach for multivariate time series," *Measurement*, vol. 191, p. 110791, Mar. 2022, doi: 10.1016/J.MEASUREMENT.2022.110791.

[14] H. Li, X. Peng, H. Zhuang, and Z. Lin, "Multiple Temporal Context Embedding Networks for Unsupervised time Series Anomaly Detection," *ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 3438–3442, May 2022, doi: 10.1109/ICASSP43922.2022.9747668.

[15] A. Putina, M. Sozio, D. Rossi, and J. M. Navarro, "Random histogram forest for unsupervised anomaly detection," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2020-November, pp. 1226–1231, Nov. 2020, doi:

10.1109/ICDM50108.2020.00154.
[16] I. Ullah, H. Hussain, I. Ali, and A. Liaquat, "Churn Prediction in Banking System using K-Means, LOF, and CBLOF," *1st Int. Conf. Electr. Commun. Comput. Eng. ICECCE 2019*, Jul. 2019, doi: 10.1109/ICECCE47252.2019.8940667.
[17] G. A. Susto, A. Beghi, and S. McLoone, "Anomaly detection through on-line isolation Forest: An application to plasma etching," pp. 89–94, Jul. 2017, doi: 10.1109/ASMC.2017.7969205.
[18] T. Huang *et al.*, "An LOF-based adaptive anomaly detection scheme for cloud computing," *Proc. - Int. Comput. Softw. Appl. Conf.*, pp. 206–211, 2013, doi: 10.1109/COMPSACW.2013.28.
[19] A. Kind, M. Stoecklin, and X. Dimitropoulos, "Histogram-based traffic anomaly detection," *IEEE Trans. Netw. Serv. Manag.*, vol. 6, no. 2, pp. 110–121, Jun. 2009, doi: 10.1109/TNSM.2009.090604.
[20] M. Xie, J. Hu, and B. Tian, "Histogram-based online anomaly detection in hierarchical wireless sensor networks," *Proc. 11th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. Trust. - 11th IEEE Int. Conf. Ubiquitous Comput. Commun. IUCC-2012*, pp. 751–759, 2012, doi: 10.1109/TRUSTCOM.2012.173.
[21] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-Based Outlier Detection in High-dimensional Data," 2008, Accessed: May 19, 2022. [Online]. Available: http://www.dbs.ifi.lmu.de
[22] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Futur. Gener. Comput. Syst.*, vol. 55, pp. 278–288, Feb. 2016, doi: 10.1016/J.FUTURE.2015.01.001.
[23] J. Liu and T. Zou, "Identifying the outlier in tunnel monitoring data: An integration model," *Comput. Commun.*, vol. 188, pp. 145–155, Apr. 2022, doi: 10.1016/J.COMCOM.2022.03.002.
[24] J. Fang, Z. Xie, H. Cheng, B. Fan, H. Xu, and P. Li, "Anomaly detection of diabetes data based on hierarchical clustering and CNN," *Procedia Comput. Sci.*, vol. 199, pp. 71–78, Jan. 2022, doi: 10.1016/J.PROCS.2022.01.010.
[25] T. Y. Kim and S. B. Cho, "Web traffic anomaly detection using C-LSTM neural networks," *Expert Syst. Appl.*, vol. 106, pp. 66–76, Sep. 2018, doi: 10.1016/J.ESWA.2018.04.004.
[26] V. Chang, L. M. T. Doan, A. Di Stefano, Z. Sun, and G. Fortino, "Digital payment fraud detection methods in digital ages and Industry 4.0," *Comput. Electr. Eng.*, vol. 100, p. 107734, May 2022, doi: 10.1016/J.COMPELECENG.2022.107734.
[27] I. T. Christou, M. Bakopoulos, T. Dimitriou, E. Amolochitis, S. Tsekeridou, and C. Dimitriadis, "Detecting fraud in online games of chance and lotteries," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13158–13169, Sep. 2011, doi: 10.1016/J.ESWA.2011.04.124.
[28] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," 2000.
[29] Z. Zhao, Yue and Nasrullah, Zain and Li, "PyOD: A Python Toolbox for Scalable Outlier Detection," *Journal of Machine Learning Research*, 2019. http://jmlr.org/papers/v20/19-011.html (accessed May 24, 2022).
[30] N. R. Prasad, S. Almanza-Garcia, and T. T. Lu, "Anomaly detection," *Comput. Mater. Contin.*, vol. 14, no. 1, pp. 1–22, 2009, doi: 10.1145/1541880.1541882.
[31] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/J.PATREC.2005.10.010.